REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
	Technical Report	-

4. TITLE AND SUBTITLE
Behaviorally Modeling Games of Strategy Using Descriptive Q-learning

5a. CONTRACT NUMBER
W911NF-09-1-0464

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Roi Ceren, Prashant Doshi, Matthew Meisel, Adam Goodie, Dan Hall

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Georgia Research Foundation, Inc.
Sponsored Research Office
University of Georgia Research Foundation Inc
Athens, GA 30602 -

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
55749-NS.3

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
Modeling human decision making in strategic problem domains is challenging with normative game theoretic approaches. Behavioral aspects of this type of decision making, such as forgetfulness or misattribution of reward, require additional parameters to capture their effect on decisions. We propose a descriptive model utilizing aspects of behavioral game theory, machine learning, and prospect theory that replicates the behavior of humans in uncertain strategic environments. We test the predictive capabilities of this model over data from 43 participants

15. SUBJECT TERMS
computational modeling, strategic games, behavioral data, reinforcement learning

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Prashant Doshi
UU	UU	UU	UU		19b. TELEPHONE NUMBER: 706-583-0827

Report Title

Behaviorally Modeling Games of Strategy Using Descriptive Q-learning

ABSTRACT

Modeling human decision making in strategic problem domains is challenging with normative game theoretic approaches. Behavioral aspects of this type of decision making, such as forgetfulness or misattribution of reward, require additional parameters to capture their effect on decisions. We propose a descriptive model utilizing aspects of behavioral game theory, machine learning, and prospect theory that replicates the behavior of humans in uncertain strategic environments. We test the predictive capabilities of this model over data from 43 participants guiding a simulated Uninhabited Aerial Vehicle (UAV) against an unknown automated opponent.

Behaviorally Modeling Games of Strategy Using Descriptive Q-learning

Roi Ceren
Department of Computer Science
University of Georgia
Athens, GA 30605
ceren@cs.uga.edu

Prashant Doshi
Department of Computer Science
University of Georgia
Athens, GA 30605
pdoshi@cs.uga.edu

Matthew Meisel
Department of Psychology
University of Georgia
Athens, GA 30605
mameisel@uga.edu

Adam Goodie
Department of Psychology
University of Georgia
Athens, GA 30605
goodie@uga.edu

Dan Hall
Department of Statistics
University of Georgia
Athens, GA 30605
danhall@uga.edu

ABSTRACT

Modeling human decision making in strategic problem domains is difficult with normative game theoretic approaches. Behavioral aspects of this type of decision making, such as forgetfulness or misattribution of reward, require additional parameters to capture their effect on decisions. We propose a descriptive model utilizing aspects of behavioral game theory, machine learning, and prospect theory that replicates the behavior of humans in uncertain strategic environments. We test the predictive capabilities of this model over data from 43 participants guiding a simulated Uninhabited Aerial Vehicle (UAV) against an unknown automated opponent.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Learning—*Parameter learning*

General Terms

Human Factors, Experimentation

Keywords

reinforcement learning, behavioral game theory, human decision making, models

1. INTRODUCTION

In strategic, uncertain environments, human decision making may not always adhere to normative decision theoretic models. When tasked with making decisions in these domains, humans do not always exhibit a clear memory of past experiences. In addition, rewards from neighboring strategies may have an impact on decisions, as humans tend to spill over rewards from one strategy to another [9]. Essentially, human decision making patterns include several cognitive biases which influence their chosen strategy.

Several behavioral game theory models exist for representing human decision making [1, 2, 7, 8]. Many of these models rely upon reinforcement learning and represent learning as the perceived reward of interaction within an environment. The application of these game theory models is limited to single-shot and repeated games which are represented in normal form.

In real-world strategic domains, the environments are largely sequential and uncertain. Reinforcement learning is well explored in these types of problem domains, for which the popular reinforcement learning technique, Q-learning, has been developed [5]. The Q-learning function determines the optimal set of strategies to maximize the total reward by analyzing immediate rewards and potential future rewards as a game progresses from state to state. Current applications of this technique apply to purely rational decision making.

This paper presents a study conducted with human subjects to observe decision making patterns. Participants in these studies were given the task of observing an unmanned aerial vehicle (UAV) navigate through a series of sectors (in a 4x4 grid) and assessing the likelihood of their UAV reaching a goal sector without being detected by an automated enemy UAV (whose location is largely unknown). The primary hypothesis of this experiment was to determine if incentivizing their assessment via proper scoring rules would improve assessment techniques. The secondary hypothesis, and the focus of this paper, was to discover if participants were learning in this environment and, if so, to model the participants' learning. While the investigation into incentives does not prove to be a significant result, we observe remarkable learning and provide an aggregate learning model.

Reinforcement learning is a convincing model for this domain. The UAV problem, while including another agent, can be modeled as a single-player game, where the participant does not model the enemy. The enemy UAV is revealed to the participant as moving in a deterministic fashion. The participant will always lose if they follow the same trajectory and are in the same state (after the same amount of moves) that caused a loss in a previous iteration of the game. Therein, the enemy is a part of the game's environment, and

need not be modeled explicitly by the participant.

The task of probability assessment in human decision making is also subject to biases [6]. When a participant states their probability assessment, it may not be equivalent to their believed probability of success. When rewards are non-deterministic, such as in gambling, there is much evidence that humans, in general, underweight or overweight their assessments at the extreme cases (near 0% or 100%) [4]. Subproportional probability weighting functions map believed probabilities to expressed assessments, which is generally not a linear mapping, as in the normative case.

While behavioral game theory, sequential reinforcement learning, and probability assessment mapping are well explored, combining them to a single model is a novel approach. We establish a formal model that attributes behavioral affects to sequential domains of uncertainty and augment assessment with a subproportional probability weighting function. We test its predictive capabilities over a data set of 43 participants. Our results indicate that this descriptive version of the Q-learning model shows significant gains over the respective normative version, as well as other baseline comparative models.

By utilizing a behavioral game theoretic model to predict human decision making, we can gain insight into the biases that humans suffer from when faced with strategic uncertainty. Models, such as our descriptive Q-learning model, are able to illustrate human learning and predict decisions that they make in strategic domains. Analyzing the parameters fit to these models measures the impact that these cognitive biases have.

2. EXPERIMENT: PROBABILITY ASSESSMENT FOR STRATEGIC DECISION MAKING

In a large study conducted with human participants, we investigate probability assessment elicited during a strategic, uncertain decision making game. We begin with a description of the game followed by a discussion of the methodology used to collect participant assessment data. We conclude this section with a description of the results generated in this study.

2.1 Study: UAV Game

To test the assessment techniques of human participants, we created a strategic game of uncertainty utilizing a graphical representation of a gameboard. In this sequential game, participants observe a UAV (hereafter *participant's UAV*) moving through a 4 x 4 sector grid from an initial sector towards a colored goal sector. Participants are given the initial location of another UAV (hereafter *enemy UAV*), but no other information about its movement or successive locations. A trial (the completion of one trajectory) is considered a "win" if the participant UAV reaches the goal sector, or a "loss" if it is caught by the enemy UAV.

Fig. 1 represents the first two sectors visited (or decision points) of a trial. The gameboard grants clairvoyance of the entire trajectory for the current trial, the initial location of the enemy, and the already traveled course.

The goal of the experiment was to gather the assessments of the overall likelihood of a trial's success from participants. Given the knowledge of the initial location of the enemy, as well as the growing knowledge of its movements based on losses, this game exemplifies a learning task.

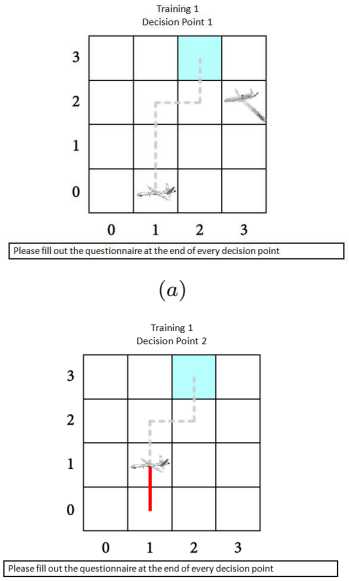


Figure 1: Two decision points of a given trial in the UAV game. The participant knows the enemy location only on the first decision point.

2.1.1 Participants

43 participants were included in this study. Participants were pulled from a pool of undergraduate students taking introductory psychology courses in our university. Participants were paid via a variety of payment mechanisms for their time. As the initial hypothesis of incentivization techniques was inconclusive, we included all participants, regardless of this effect, in this paper.

2.1.2 Methodology

Participants play 20 total trials of the game. Two initial phases, representing the training phases of the game, consist of 5 trials each. At the end of each of these sets, the participant undergoes an intervention, in which the proctor of the experiment highlights participant assessments which are too high or too low.

At each decision point, the participant is required to fill out a questionnaire. In the questionnaire, the participant notes the direction the UAV will move and their estimation for the probability that the participant UAV will, without being caught, arrive in the next sector and the eventual goal sector. After filling out the questionnaire, the participant may move onto the next slide of the game.

2.1.3 Results

Participant data was broken up into two discrete data sets: trials resulting in wins and those resulting in losses. We analyzed the data for trends within the trial (as the UAV approached the goal sector) and between trials (as participants became more familiar with the game). We expect, as a trial progresses, that a participant will assess higher likelihoods of success as they approach the goal sector. Additionally,

as the game progresses, the participant should become more confident in their assessments.

Table 1: Slope analysis of results

Trend	Estimate	P-value
intercept	0.3315	<0.0001
slope within trial	0.02053	0.0016
slope across trials	-0.00486	<0.0001

(*a*) Losses

Trend	Estimate	P-value
intercept	0.5392	<0.0001
slope within trial	0.05395	<0.0001
slope across trials	-0.00129	<0.0001

(*b*) Wins

Table 1 above annotates the results of running a generalized linear mixed effect regression analysis over our data with random intercept and slope at the decision point and trial level. Our results indicate that the estimates given for each point is significant.

When considering assessments as a trajectory progresses, participants generally increase their assessments as they approach the goal sector. The rate by which a participant's stated probability increases for winning trajectories is greater than losses. This is to be expected, as participants will become more familiar with the possible movement of the enemy, they will become better at predicting eventual losses.

As participants complete trials, the slope of the change in elicited probabilities decreases significantly. This decrease in slope indicates that participants are not changing their probability assessments as much as they were in previous trials, representing a general increase in confidence of the participant's guesses for both wins and losses. The ideal case is that, as participants learn how the enemy is moving, their slope across trials will approach 0.

With the clear trends towards generally increasing assessments as trials progress and the relative growth of confidence as participants complete trials, these results indicate a strong justification for the application of a learning model.

3. DESCRIPTIVE MODEL FOR REINFORCEMENT LEARNING

Our model is an extension of the popular reinforcement learning algorithm known as Q-learning. By attributing concepts derived from behavioral game theory to Q-learning, we establish a novel framework for descriptive reinforcement learning. Additionally, borrowing from concepts in prospect theory creates a better mapping of true beliefs to expressed probabilities.

3.1 Normative Q-learning

Q-learning is a popular machine learning model for representing learning in sequential domains. It characterizes the reinforcement learning problem as a conjunction of previous information and future rewards, decayed by a discount parameter, γ. Q-learning is an algorithm that exemplifies *exploration* vs. *exploitation*, which prefers possible future payoffs or previously learned payoffs, respectively [5]. This decision is mediated by the learning parameter, α. Equation 1 shows the standard Q-learning function.

$$Q(s,a) = Q(s,a) + \alpha(r(s) + \gamma max_{a'} Q(s',a') - Q(s,a)) \quad (1)$$

This function serves as a powerful mechanism to model learning with long-term optimality. However, it does not exemplify the behavioral aspects of human decision making. With the concepts derived from behavioral game theory, we can apply descriptive parameters to the Q-learning function.

3.2 Behavioral Q-learning

The inspiration for the descriptive model is derived from behavioral game theory. Several game theoreticians [2, 9, 3] have investigated human biases as associated with problems of decision making. Their investigations are uniquely in the context of single shot and repeated games.

3.2.1 Behavioral Reinforcement Learning

Game theory seeks to analyze and explain the mechanisms by which decisions are made [1]. Assuming that participants understand the game, the environment, and make decisions in a purely rational manner, applicable game theoretic models will be able to predict the behavior of a human. This is rarely the case in reality, however. Cognitive biases plague the human decision making process, leading to seemingly subrational decisions. Behavioral game theory models learning with these biases in consideration.

Several models exist that attempt to express learning within decision making domains. The reinforcement learning algorithm portrays learning as a function of interaction with an environment and the immediate rewards. As an individual moves through the world, it experiences stimuli that it attributes to doing a particular action. Algorithmically, the reinforcement learning algorithm can be characterized as:

$$A_c(t) = A_c(t-1) + r \quad (2)$$

The attraction to doing a strategy c at time step t is the previous attraction to doing strategy c and its immediate reward. An attraction may be implemented in many ways, but it is essentially a concept representing the desirability of taking a particular action.

Insights from behavioral game theory have provided parameters that better explains the irrational behavior that arises in human decision making [2]. Such concepts include forgetfulness (the event of previous information degrading in effect on future decisions) and spillover (the phenomenon of humans attributing rewards to neighboring strategies). Behavioral reinforcement learning can be expressed as:

$$A_c(t) = \phi A_c(t-1) + (1-\epsilon)r \quad (3)$$

$$A_n(t) = \phi A_n(t-1) + (\epsilon)r \quad (4)$$

ϕ represents the forgetfulness parameter, ϵ represents the spillover parameter, and A_n represents the attraction to strategy n, which is a neighboring strategy to c. Both parameters are bounded between 0 and 1.

Forgetfulness in the context of our domain would imply that the experience from a previous trial has a diminished effect on current experiences. Spillover generally involves the misattribution (or "generalization") of rewards to neighboring strategies. An illustrative example is that of the roulette

player who places a large bet on a particular number, only to have it land on a nearby number [9]. The player may have his guess confirmed, since the ball was near their bet, regardless that they lost the bet.

The implementation of the spillover parameter can be conceptualized in a few different ways for our UAV domain. Neighboring strategies can be viewed as nearby sectors, directly adjacent to the sector arrived at. Since the enemy moves in a deterministic pattern, the amount of moves that have transpired is directly related to the current location of the enemy. With this in mind, spillover can also occur between these *time steps*. Figure 2 exemplifies the various models that could represent spillover in this domain.

With Camerer et al.'s introduction of behavioral parameters in human decision making, we now introduce our Q-learning function as inspired by these concepts.

3.2.2 Modified Q-learning Function

$$Q(s,a) = \phi Q(s,a) + \alpha((1-\epsilon)r(s) + \gamma max_{a'} Q(s',a') - \phi Q(s,a)) \quad (5)$$

$$Q(s_n,a) = \phi Q(s_n,a) + \alpha((\epsilon)r(s) - \phi Q(s_n,a)) \quad (6)$$

ϕ, as with its behavioral game theory counterpart, represents the forgetfulness parameter, which decays the value of previous information associated with that state (in our case, waypoint sector). α mediates between exploration or exploitation, and additionally decays future payoffs to better value current information about the state as it approaches 1. If ϵ is greater than 0, the neighboring states (notated as s_n and includes all sectors that are 1 move away) gain a fraction of the reward observed [2] [7].

The future payoff calculation in the Q-learning function is of questionable application to our problem domain, however. In essence, $max_{a'}$ assumes that the future state-action pairs will be the optimal choice. Participants in our problem domain do not select the movements of the UAV, however. With clairvoyance over the trajectory that the UAV will travel, participants are likely to base their assessment on the path revealed to them.

$$Q(s,a) = \phi Q(s,a) + \alpha((1-\epsilon)r(s) + \gamma Q(s',\pi(s')) - \phi Q(s,a)) \quad (7)$$

Equation 7 alters the future payoff function to represent a *next state payoff* from the next sector, determined from the path revealed to the participant. $\pi(s')$ represents the action determined from being in state s', which, in our case, is the next sector in the trajectory for the given trial.

4. PERFORMANCE EVALUATION

The data collected from the 43 participants from this study were broken up into 5 folds, with 8-9 participants per fold. Utilizing the Nelder-Mead method[1], parameters are trained over 4 folds and then, to test the predictive capabilities of the model, tested over the remaining fold. For a baseline comparison, fits were generated for the normative model[2] and compared with the descriptive model, along with the random model[3] and pathological cases[4].

Prior to calculating the fit of the descriptive model, we must convert the calculated Q-value generated by the descriptive model to a probability assessment that will be compared to the participant data. Q-values for all states are initialized to 0, with a Q-value of 1 being allocated to the goal state and -1 for all loss states. Q-values approaching -1, then, represent a path likely to lead to a loss, whereas those approaching 1 indicate a possible win from that path. To convert these values to assessments, then, involves normalizing the Q-value between 0 and 1. The resulting conversion is then used as the Q-learning function's assessment.

Fits were generated by taking the squared distance between the participant's stated probability and the model's generated probability at each decision point in the game. The model was subjected to a simulation of the game, where it was presented with the same trajectories and experienced the same outcomes as participants. At each point where a Q-value was updated (following a simulation of a leg of a trajectory), the distance between all participants' probability assessments and the estimated Q-value were squared, aggregated, and added to the total fit.

Table 2: Spillover fits

Fig. 2.b	Fig. 2.c	Fig. 2.d	Fig. 2.e
415.534	415.924	416.122	409.254

In generating the results, we found the best fits to adhere to Figure 2.e, annotated in Table 2. This indicates that participants considered negative and positive payouts to be irrespective of the decision point. Essentially, if a participant were to lose in sector [1,2] in the 3rd decision point, they would evaluate sector [1,2] negatively in the 2nd and 4th decision point as well, while also avoiding neighboring sectors ([1,1], [1,3], and [2,2]) in those time steps as well.

Table 3: Descriptive model: parameters and fits

α	ϕ	ϵ	Fit
0.819	0.591	0.537	409.254

Table 3 annotates the results from optimizing our Q-learning function utilizing the Nelder-Mead method. Our comparative analysis between models is described in Table 4. The descriptive model outperformed the normative model with $p < 0.01$. Additionally, the descriptive model had a better test fit than the random and pathological models.

4.1 Improving Model Predictions

Although our results are significant, improvements can be made to the predictive capabilities of our model. Humans not only exhibit cognitive biases in generating their probabilities, but they additionally misrepresent those probabilities [6]. By including a theoretically sound probability weighting function, we improve our descriptive model by replicating this behavior.

[1]The Nelder-Mead method is a downhill simplex method for minimizing an objective function

[2]The normative model does not include any descriptive parameters. $\phi = 1$ and $\epsilon = 0$, while α is still trained.

[3]The probability estimations are completely random for each decision point within a trial.

[4]Pathological cases include the categorical optimist and pessimist (who always guess 100% and 0%, respectively)

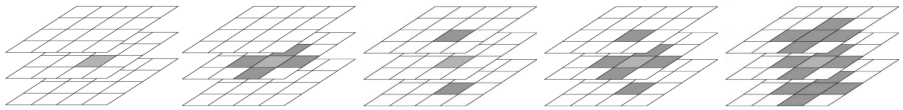


Figure 2: (a) No spillover, (b) local spillover, (c) time step spillover, (d) fractional time step and location spillover, (e) full time step and location spillover.

Table 4: Fit comparison for all models

Descriptive	Normative	Random	Optimist	Pessimist
409.254	416.409	891.182	1052.723	1971.333

4.1.1 Probability Weighting

Prospect theory notes that the weight given to probability assertions and the associated payoff values are usually not linear. That is, humans tend to under- or over-weight probability assessments in domains of chance. In our domain, participants are queried with their assessment for the overall success of their current trial as it progresses, which is subject to non-linear assessment mappings. To this end, we included a subproportional function in the mapping of Q-values to probability assessments [6].

$$w(p) = exp(-(-ln(p))^{\beta}) \tag{8}$$

Equation 8 defines the subproportional function for a given probability, p. Between 0 and 1, the exponent β causes the curve to be inverse sigmoidal. This indicates that probabilities are overweighted when low and underweighted when high. Inversely, if β is above 1, the curve becomes sigmoidal. At 1, the curve is linear, which is the normative case. Figure 3 illustrates the curves generated from example values.

4.1.2 Results

We ran the same simulation from the original descriptive model on the probability weighting descriptive model. As in the original model, we also compared the augmented model with the 43 participants from the UAV study, aggregating fits by squaring the distance from the probability weighting descriptive model's Q-values to the participants' probability assessments.

Table 5: Descriptive model: parameters and fits

α	ϕ	ϵ	β	**Fit**
0.677	0.378	0.273	1.573	401.36

Including Prelec's probability weighting function improved the performance of the descriptive model. Table 5 describes the averages for the parameters across folds and the fit generated by the model. Both α and ϕ decreased as a result of the inclusion.

Table 6: Comparative Fits

Descriptive (Weighted)	Descriptive (Unweighted)
401.36	409.254

Table 6 shows a side-by-side comparison of the descriptive model's fit both with and without the probability weighting function. A two-tailed T-test of the distance between the each model version's generated probability resulted in a significant p-value of less than 0.01. Since the weighted model is a significant improvement over the unweighted model, it is, transitively, an improvement over the normative model as well.

5. ANALYSIS

5.1 Parameters

Analysis of the test fits for the descriptive model illuminated some behaviors of human participants in sequential strategic games. The first observation we made is that the higher value for β is representative of a decision making pattern that may be characteristic of win-or-lose strategic games. Traditionally, in betting games, participants tend to avoid extreme estimations [6]. However, in the unknown environment of our particular domain, a cursory glance at the raw data indicates a predilection towards extreme probability assessments, which our model corroborates.

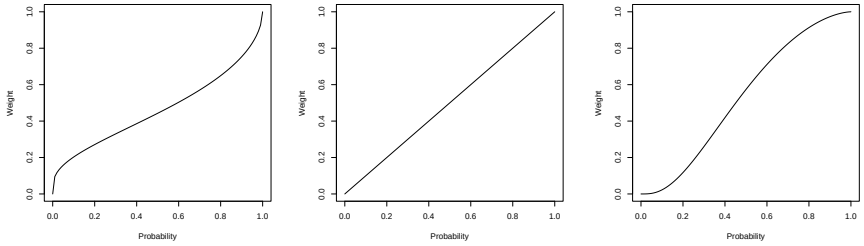
The results also indicate a higher preference for exploitation of knowledge in our domain. ϕ values converged, on average, near 0.5, with slightly higher α values. A ϕ value towards 0.38 would indicate that participants' previous knowledge is deteriorating at a rate of about a third of the reward from the last time the state was visited. α tuned around 0.677 would indicate a higher rate of exploration as participants move through the game. That is, participants are valuing new information at 68% of its actual reward.

The observation of the ϵ parameter bears discussion, as well. A spillover rate of 27% is relatively high in comparison to other implementations of this parameter in reinforcement learning [2]. This would indicate that participants were attributing around a quarter of the received reward for a sector to its neighboring sectors.

5.2 Projected probabilities

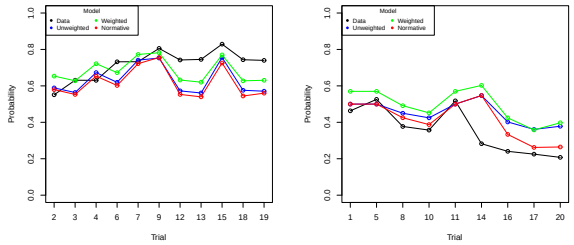
As with our cursory analysis of the data received from participants, plots of the models' probability estimations were categorized by wins and losses when compared with the estimates made by participants.

Figure 4 plots the average probabilities for trials generated from the various models (descriptive with weighting, descriptive without weighting, and the normative model) and the data. Figure 4.a shows a relatively similar curve between the models and the data, with the descriptive model with



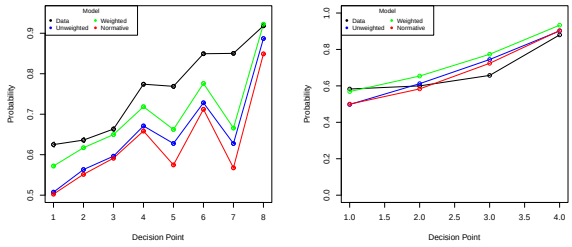
(a) (b) (c)

Figure 3: (a) $\beta = 0.56$, (b) $\beta = 1$ (linear), (c) $\beta = 1.6$

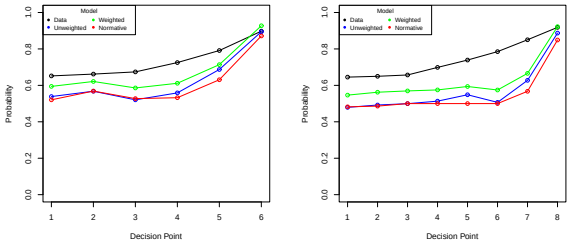


(a) Wins (b) Losses

Figure 4: Trial averages



(a) All points (b) 4 points



(c) 6 points (d) 8 points

Figure 5: Decision point averages (wins)

weighting being the closest in overall distance. For figure 4.b, the shape is also similar to the data, but the descriptive model with weighting is no longer the closest. As we'll see with later plot analysis, the models are less accurate on the trials that result in a loss, indicative of a different type of learning and probability assessment in those cases.

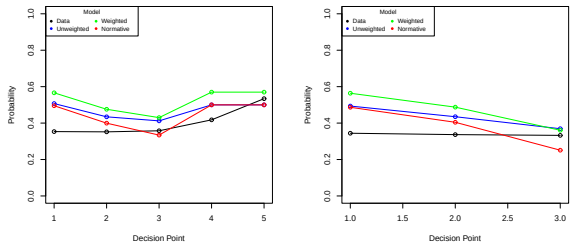
Figure 5 shows the plots for the averages of probability assessments for individual decision points made by participants and generated by the models for trials that resulted in a win. Trials that result in wins can be categorized into 3 different trajectory lengths. If the participant's UAV eventually reaches the goal sector, it will do so in 4, 6, or 8 moves. Figure 5.a shows the overall plot for averages of decision points regardless of the trial type. While the overall fit for the descriptive model with weighting is the closest, the plot has a strange shape. This is due to the different amount of data points for trials of different lengths (e.g. there are only 3 trials of length 8, but there are 11 total trials that result in a win) and the different types of behavior in the various trial lengths. Figures5.b, 5.c, and 5.d show the underlying behavior for trials of each length, with the descriptive model with weighting outperforming the other models in each case.

Figure 6 shows the plots for averages of probability assessments for the data and models over trials that consist of losses. These trials break down into 3 and 5 point trials and are categorized accordingly. As with the plot for the loss trial averages, the models tend to perform worse on decision point averages for loss trials. Participants, on average, start with much lower assessments than with trials that result in a win. This indicates that participants are better at identifying eventual losses and retain their pessimism as trials progress. The models, on the other hand, become progressively more pessimistic. The data for the 5 point trial is completely flat as there is only one trial that is 5 points in length (that results in a loss) and the model is not able to acquire enough information to give an accurate assessment.

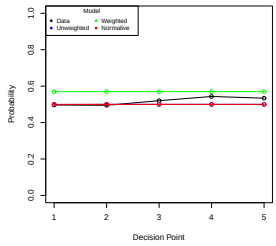
5.3 Discussion

The results of the fitting of this model are illuminating. They are indicative of the relative power of behavioral game theoretic parameters in a sequential learning model. The addition of a probability weighting curve further improved our results.

Though the analysis on reinforcement learning in this domain indicates a significant gain from the inclusion of behavioral parameters, other competing learning models can be compared as a baseline for the effectiveness of reinforcement learning in this domain. Several behavioral approaches to belief-based learning may be applicable to the sequential strategic game utilized in this paper. Camerer et al. have proposed alternative models to reinforcement learning in be-



(*a*) All points (*b*) 3 points



(*c*) 5 points

Figure 6: Decision point averages (losses)

havioral game theory that may beg further investigation.

6. RELATED WORK

Several other models exist that seek to express descriptive learning in human decision making domains. Besides reinforcement learning, belief learning, experience-weighted attraction learning, imitation, and direction learning also represent other approaches to behavioral game theory [1]. Belief learning represents learning as a process of basing future considerations on observed behavior in the last round [3]. In our domain, it is possible for participants to consider their rewards as dependant on the movement of the enemy, but, considering the lack of information associated with the enemy, it is likely that their wins and losses are modeled as an aspect of the environment.

Erev and Roth also investigate descriptive reinforcement learning, but it is examined in repeated stage games, not the sequential domain [8]. Many of the applications of our model are present in their work, but the concept of uncertainty and generalizations of strategy are not implemented in their analysis.

Our work extends observations from Camerer et al.'s *Experience-Weighted Attraction* model, though it has similar shortcomings as the Erev and Roth model [2]. This model is contextual to stage games, as opposed to the sequential environment of the UAV and other strategic problems. Additionally, the applicability of the law of simulated effect[5] is less pronounced in our model, as the payouts for foregone strategies are unknown.

[5]The law of simulated effect states that foregone strategies that are known to have produced better results if chosen will have a higher attraction in subsequent games.

Acknowledgments

This work was supported by a grant from the Army RDECOM, grant #W911NF-09-1-0464, to Prashant Doshi, Adam Goodie and Dan Hall. We thank Xia Qu for providing help and support during the conduct of this research.

7. REFERENCES

[1] C. Camerer. *Behavioral Game Theory*. Princeton University Press, Princeton, New Jersey, 2003.

[2] C. Camerer and T. Ho. Experience-weighted attraction learning in normal form games. *Econometrica*, 26(4):827–874, July 1999.

[3] A. Cournot. *Recherches sur les principes mathématiques de la théorie des richesses*. Haffner, London, 1960. Translated by N. Bacon as *Researches in the Mathematical Principles of the Theory of Wealth*.

[4] R. Gonzales and G. Wu. On the shape of the probability weighting function. *Cognitive Psychology*, 38(1):129–166, February 1999.

[5] L. Kaelbling, M. Littman, and A. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, May 1996.

[6] D. Prelec. The probability weighting function. *Econometrica*, 4(3):497–527, May 1998.

[7] A. Roth and I. Erev. Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior*, 8:164–212, 1995.

[8] A. Roth and I. Erev. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *The American Economic Review*, 88(4):848–881, 1998.

[9] W. Wagenaar. *Paradoxes of Gambling Behavior*. Lawrence Erlbaum, Mahwah, New Jersey, 1984.